Supplementary material for Class Semantics-based Attention for Action Detection

Deepak Sridhar*† Niamul Quader* Srikanth Muralidharan Yaoxin Li Peng Dai Juwei Lu
Huawei Noah's Ark Lab, Canada
{deepak.sridhar1,niamul.quader1,srikanth.muralidharan,yaoxin.li,peng.dai,juwei.lu}@huawei.com

1. Ablation Experiments

Sensitivity to hyper-parameters and model sizes: To provide more detailed insights into the impact of CSA, we experiment with different hyperparameters of the BMN-CSA model including number of samples used to construct BM features (N) in BMN [1] and the output number of channels of the encoder layer (C_{out}) which heavily affects the overall model size in terms of number of parameters. Table 1 summarizes these results, where we compare BMN-CSA against baseline BMN model. We observe that BMN-CSA obtains consistent improvement over baseline with respect to BMN across different hyperparameter settings, and also achieves state-of-the-art performance of **35.75**. We also note that the parameter complexity increases considerably with the C_{out} hyperparameter ($\mathcal{O}(C_{out}^2)$). As demonstrated in Table 1, incorporating our CSA on BMN with different number of model parameters (or different C_{out} values) provides consistent performance gains over baseline BMN, suggesting that CSA will likely be useful for action localization networks with varying model sizes.

Alternate aggregation strategies for CSA: First, we experiment alternate channel/temporal CSA branch aggregation strategies. Specifically, we compare elementwise multiplication based aggregation and adaptive (learned weighted) addition based aggregation against the current BMN-CSA model. The learned weights corresponding to the adaptive aggregation method in Table 2 is - Channel attention branch weight: 0.55, Temporal attention branch weight: 0.45, it weighs channel attention slightly higher than the temporal attention. Based on the results shown in Table. 2 we observe that BMN-CSA outperforms these alternate aggregations necessitating the learned concatenation based aggregation we perform in BMN-CSA model.

Applying CSA at a different layer: Second, we explore alternate model design strategies, where we apply CSA at locatons different from our BMN-CSA model. Before discussing our alternate designs, we provide a brief overview

N	Method	$C_{out} = 128$	$C_{out} = 256$	$C_{out} = 320$
8	BMN	35.00	34.89	34.89
	BMN-CSA	**35.26** (+0.26)	**35.35**(+0.46)	**35.36**(+0.47)
16	BMN	35.03	34.90	34.66
	BMN-CSA	**35.75**(+0.72)	**35.39**(+0.49)	**35.53**(+0.87)
24	BMN	34.89	34.91	34.86
	BMN-CSA	**35.45**(+0.56)	**35.6**(+0.69)	**35.43**(+0.57)
32	BMN	34.72	34.88	34.88
	BMN-CSA	**35.34**(+0.62)	**35.43**(+0.55)	**35.44**(+0.56)

Table 1. Validation mAP obtained on ActivityNet dataset using the baseline BMN as well as BMN-CSA by tuning the following hyperparameters: 1. Number of samples (N) considered for BM feature construction and 2. Output number of channels C_{out} of channel branch of CSA (this also corresponds to number of channels of localizer input). Note that we obtain a new state of the art of **35.75**, further higher than the mAP reported in the paper from this hyperparameter tuning experiment.

CSA Aggregation	mAP
Multiplication	34.9
Adaptive addition	35.14
BMN-CSA	**35.75**

Table 2. Alternate CSA branch aggregation strategies compared against current BMN-CSA aggregation. Comparisons made on Activitynet validation set.

CSA Application	mAP
PEM-CSA	35.07
TEM-CSA	34.97
TEM-PEM CSA	35.01
BMN-CSA	**35.75**

Table 3. Alternate CSA design strategies, where we apply CSA at locations different from BMN-CSA. Specifically, we compare alternate designs such as applying CSA on TEM module (TEM-CSA), on PEM module (PEM-CSA), on both TEM and PEM modules (TEM-PEM CSA).

*denotes equal contribution
†Corresponding author

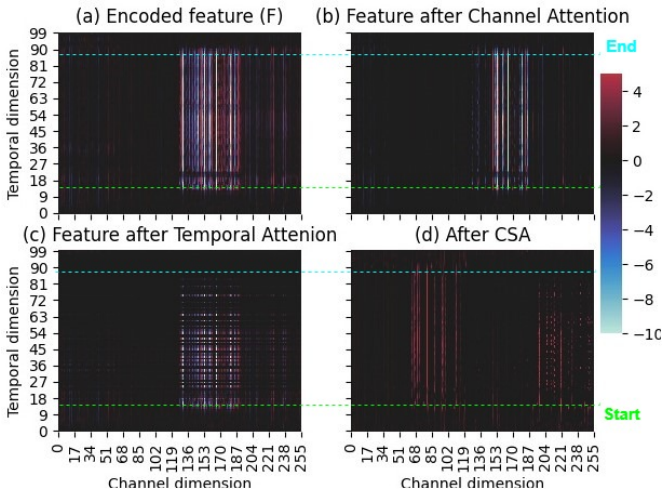

Figure 1. Comparison of the activation maps - (a) F, (b) F_{A_T}, (c) F_{C_T}, and (d) the CSA modified feature map. Note that since start and end are semantically opposite, if high attention weights (and corresponding high-magnitude activation maps values) are seen near start-point regions, conversely, we would expect low attention weights (and corresponding low-magnitude activation maps values) near end-point regions. This is observed in (c).

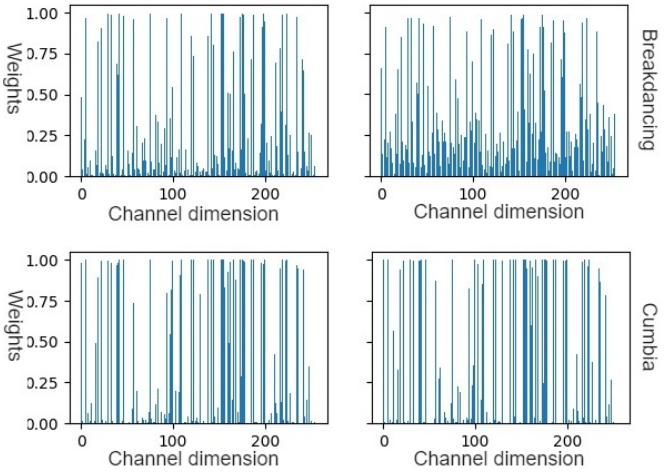

Figure 2. Channel attention weight profiles on four separate videos show more similar profiles within same classes (left vs. right) compared to profiles from different classes (top vs. bottom).

of BMN model architecture. BMN's encoder sub-network learns feature representation suitable for action detection. Its localizer sub-network jointly predicts proposal boundary probability map using "Temporal Evaluation Module (TEM)", and confidence score for proposals using "Proposal Evaluation Module (PEM)".

In our alternate design, we consider applying CSA at TEM, which we refer to as TEM-CSA, PEM-CSA to refer to applying CSA at the beginning of PE module, and TEM-PEM CSA to refer to applying CSA at both these layers. Table. 3 tabulates mAP obtained on ActivityNet dataset using these alternate network designs. We observe that BMN-CSA outperforms alternate designed choices we considered in this section.

Effect of applying CSA at multiple locations: In the paper, we showed experiments to demonstrate the effect of applying CSA at different locations within the localization encoder module. Here, we also conduct another experiment on applying CSA to both at middle and last of the encoder module. Table 4 shows the results of this experiment on Thumos dataset. We see that applying CSA at both middle and the end of encoder module significantly decreases the accuracy compared to applying it at the end only (though it still remains better than baseline BMN). For ActivityNet, we observe a similar result where BMN-Middle+Last CSA obtains 35.36 mAP as compared to applying CSA at the end of encoder module of BMN (35.43 mAP). We conjecture that because the input to CSA is the same for attention applied to both locations, it does not learn significantly different attention weights for different locations rather it increases the size of the attention modules in terms of number of learnable parameters and makes it more difficult to train. Hence, applying CSA at multiple locations will likely not provide any further performance gain compared to applying it at the end of encoder module only.

Module	0.3	0.4	0.5	0.6	0.7	mAP
Middle-CSA	63.8	57.1	47.7	36.9	25.3	46.2
Last-CSA	**64.4**	**58.0**	**49.2**	**38.2**	**27.8**	**47.5**
Middle+Last-CSA	61.1	53.4	43.6	32.9	21.0	42.4

Table 4. Ablation on applying the CSA module at multiple locations of the encoder module on Thumos'14 dataset

1.1. Visualization of attention maps

We plot the activation maps and the learned attention weights in order to visually locate the regions that are activated by our proposed attention mechanism. Fig. 1 shows the comparison of the activation maps before and after applying attention. We observe that the attention weights are semantically opposite for the start and end boundary locations, thus making it more discriminative for action proposal generation. Fig. 2 shows the attention profile for two different classes - Breakdancing and Cumbia. We notice that the attention profile along the channel dimension varies across different classes while being similar within same the class (dense profile for two different breakdancing videos vs sparse profile for two different cumbia videos).

References

[1] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3889–3898, 2019. 1